Grundfragen der digitalen Langzeitarchivierung für den edoc-Server

Susanne Dobratz | dobratz@cms.hu-berlin.de

In diesem Artikel wird das LZA-Konzept „edoc-Server" vorgestellt. Dabei werden verschiedene Methoden beleuchtet, angefangen bei dem Einsatz archivierungsfähiger Dateiformate, bis hin zur Herstellung technischer Redundanz über RAID-Verfahren.

Digital Information lasts forever or five years – whichever comes first.
Jeff Rothenberg [1]

Der edoc-Server der Humboldt-Universität zu Berlin

Seit 1998 ist der edoc-Server an der Humboldt-Universität in Betrieb. In den vergangenen 11 Jahren haben wir als Betreiber des Servers so einige Technologiewechsel miterlebt. Wir haben die Hardware des edoc-Servers mehrfach erneuert und entsprechend dazu neue Betriebssystemversionen des Sun Solaris-Betriebssystems installiert. Wir haben den Kern unseres edoc-Servers, die Web-Anwendungen, ausgebaut. Vor allem aber gab es seitdem einige Versionswechsel des Textverarbeitungssystems Microsoft Word, mit dem der Großteil der Autoren seine Publikationen erstellt. Diese haben dazu geführt, dass die Dokumentvorlagen, besonders die für die Dissertationen, mehrfach angepasst werden mussten. Auch das Publikationsformat PDF hat seit 1998 mehrere Versionswechsel erlebt.

Wagt man unter diesen Bedingungen einen Blick in die Zukunft, so ist eines sicher:Technologiewechsel finden im IT-Bereich sowohl bei der Hardware als auch bei der Software relativ schnell statt. Man geht davon aus, dass alle 2–3 Jahre eine Umstellung einzelner Komponenten erfolgen muss. Diesen stetigen Wandel an Technologie zu bewältigen und den edoc-Server als stabile Dienstleistung bereitzustellen stellt eine große Herausforderung dar. Denn das Ziel ist es, den edoc-Server als zentrale Publikationsplattform für die Universität auch in den nächsten Jahrzehnten zu betreiben.

So heißt es in den Leitlinien des edoc-Servers: „Der Dokumenten- und Publikationsserver bietet allen Angehörigen der Humboldt-Universität die organisatorischen und technischen Rahmenbedingungen zur elektronischen Publikation wissenschaftlicher Dokumente. Im Rahmen dieses Gemeinschaftsangebotes von Computer- und Medienservice und Universitätsbibliothek der Humboldt-Universität werden wissenschaftliche Dokumente von hoher Relevanz unter Einhaltung von Qualitätsstandards im Internet für Forschung und Lehre bereitgestellt. Die elektronischen Dokumente erhalten dauerhafte Adressen und werden über nationale und internationale Bibliothekskataloge, Suchmaschinen sowie andere Nachweisinstrumente erschlossen. Der Dokumenten- und Publikationsserver bietet durch besondere Maßnahmen wie digitale Signaturen und Zeitstempel einen Schutz gegen Verfälschungen. Darüber hinaus wird eine Langzeitarchivierung der elektronischen Dokumente gewährleistet." [2]

Die Problematik der Langzeitarchivierung wurde beim Aufbau des Dokumentenservers der HU von vornherein als ein wichtiges Thema angesehen, und die dazu erforderlichen Maßnahmen wurden und werden schrittweise umgesetzt.

Was versteht man unter digitaler Langzeitarchivierung?

Das Ziel der Archivierung digitaler Dokumente ist es, diese für eine zukünftige Nutzung so aufzubewahren, dass kommende Generationen darauf zugreifen, sie benutzen und Zugang zu dem darin kodierten intellektuellen und kulturellen Inhalt finden können. Idealerweise würde man davon ausgehen, dass jedes digitale Dokument in seiner heutigen Form und Funktionsweise verfügbar und benutzbar gehalten wird. Allerdings ist dieses Ideal kaum zu realisieren, da die Technologien sich sehr schnell und mitunter auch sehr stark verändern. Deshalb gehen heutige Strategien zur digitalen Archivierung davon aus, dass es unmöglich ist, zukünftige Technologien und Anwendungsszenarien soweit im Voraus zu bestimmen. Daher ist es im Allgemeinen das vorrangige Ziel, zum einen den Zugriff auf die digitalen Dokumente sicher zu stellen und zum anderen deren Lesbarkeit.

Wenn man z. B. ein eigenes Dokument heute in digitaler Form auf einem universitären Dokumentenserver bereitstellt, möchte man natürlich auch, dass die Enkel und Urenkel dieses Werk noch lesen und benutzen können. Dies gelingt allerdings nur, wenn die Universität dafür sorgt, dass entsprechende Maßnahmen für die Langzeitarchivierung der digitalen Publikation getroffen werden. Digitale Langzeitarchivierung beginnt beim Produktionsprozess. Als Autor schreibt man meist mit dem am besten verfügbaren Textverarbeitungsprogramm, etwa Microsoft Office oder auch OpenOffice. Dabei stellt sich die Frage, ob diese Dateien in drei oder fünf Jahren noch in dem Nachfolgeprogramm geöffnet werden können. Viele Computernutzer haben die Programme, mit denen die Dokumente vor einigen Jahren geschrieben worden sind, bereits heute nicht mehr auf ihrem Arbeitsplatzrechner verfügbar. Dokumente müssen deshalb in einem standardisierten Dateiformat publiziert werden, etwa in PDF oder XML. Aber auch diese Formate werden möglicherweise im Laufe der Zeit unbenutzbar, sodass die Anwendung einer Archivierungsstrategie ratsam ist, die eine regelmäßige Migration der Dokumente in aktuelle Dokumentformate vorsieht. An dieser Stelle fällt der Blick auf das Speichermedium, das benutzt wurde. Auch die DVD oder magnetische Medien wie Festplatten altern physikalisch. Demnach muss die Speicherhardware auch regelmäßig überprüft und erneuert werden. Die Identifizierung des zum betrachteten Dokument gehörenden Bitstroms und die Identifikation und Interpretation des Dateiformates spielen eine wichtige Rolle für dessen spätere Benutzbarkeit. Dazu kommt die Bereitstellung des Daten- oder Zusatzmaterials, der Statistiken oder der Videos, die zum Inhalt heutiger wissenschaftlicher Arbeiten gehören. Welches aber sind die Standards, die wir heute schon einsetzen können, um zukünftigen Generationen den Zugang und die Benutzbarkeit zu unserem digitalen Kulturerbe zu garantieren? Der schnelle Technologiewechsel unserer Zeit macht es uns unmöglich, diese Art von Garantie zu geben. Im Gegenteil, wir werden permanent damit beschäftigt sein, diese Technologiewechsel zu beobachten und angemessen darauf zu reagieren.

> „Langzeitarchivierung ist nicht die Abgabe einer Garantieerklärung über fünf oder fünfzig Jahre, sondern die verantwortliche Entwicklung von Strategien, die den beständigen, vom Informationsmarkt verursachten Wandel bewältigen können." [3]

Dabei spielen sowohl organisatorische als auch technische Maßnahmen eine große Rolle, wie sie u. a. im „Kriterienkatalog vertrauenswürdige digitale Langzeitarchive" [4] formuliert werden. Sie schließen sowohl die Migration als auch die Emulation als Erhaltungsmethode ein.

Unter Migration wird eine Menge organisierter Abläufe verstanden, die den regelmäßigen Transfer digitalen Materials von einer Hardware/Software-Konfiguration in eine andere realisiert oder von einer Computer-Technologie-Generation in eine nachfolgende Generation. Dagegen versteht man unter Emulation die Methode, ein digitales Dokument (mit seiner originalen Kodierung) unter Nutzung einer speziellen Software (eines Emulators) in seiner ursprünglichen Softwareumgebung nutzbar zu machen.

Organisatorische Rahmenbedingungen

Langfristig soll der edoc-Server zu einem vertrauenswürdigen Langzeitarchiv ausgebaut werden.

Digitale Langzeitarchive haben den Erhalt der Informationen über lange Zeiträume hinweg zum Ziel. Deshalb ergreifen sie sowohl organisatorische als auch technische Maßnahmen, um den Bedrohungen, denen die digitalen Objekte ausgesetzt sind, entgegenzuwirken. Dazu gehört die Nutzung von Standards in Bezug auf die Hardware, die Speicherformate der digitalen Objekte, die Beschreibung mit angemessenen Metadaten sowie der Einsatz standardisierter Verfahrensweisen bei der Migration oder der Emulation digitaler Objekte. Vertrauenswürdige digitale Langzeitarchive operieren nach ihren Zielen und Spezifikationen, vgl. [4].

Bis dies realisiert ist, ist die Humboldt-Universität zu Berlin eine Kooperation mit der Deutschen Nationalbibliothek (DNB) eingegangen. Die über den edoc-Server publizierten digitalen Dissertationen und Habilitationsschriften werden in das DNB-eigene digitale Langzeitarchiv übernommen.

Einige Eigenschaften eines vertrauenswürdigen digitalen Langzeitarchivs besitzt der edoc-Server aber bereits schon heute? Dazu gehört an erster Stelle die Garantie eines zuverlässigen und organisatorisch gesicherten Betriebes. Diese Garantie hat die Humboldt-Universität zu Berlin dadurch gegeben, dass sie den edoc-Server als festen Bestandteil in das Dienstleistungsangebot von Computer- und Medienservice und Universitätsbibliothek eingegliedert und den Betrieb durch zugehöriges Personal abgesichert hat.

An zweiter Stelle wurden für den edoc-Server Leitlinien definiert, die beschreiben, welche Publikationen unter welchen Bedingungen über den edoc-Server veröffentlicht werden und unter welchen Bedingungen Garantien für eine zukünftige Lesbarkeit und Benutzbarkeit der Dokumente gegeben werden.

Drittens werden der Zugang zum und der Zugriff auf den edoc-Server über diverse organisatorische und technische

Maßnahmen kontrolliert und gesichert. So steht der Server selbst im Rechnerraum des Computer- und Medienservice und ist damit in die Sicherheits-Infrastruktur des Computer- und Medienservice eingebunden. Durch den Einsatz von digitalen Signaturen wird sichergestellt, dass die Authentizität und Integrität der Dokumente gewahrt wird und jede nachträgliche Änderung der Autorenschaft, des Veröffentlichungsdatums oder gar des Inhaltes bemerkt wird. Des Weiteren wird durch die edoc-spezifische Backupstrategie dem Verlust von Dokumenten vorgebeugt.

Ein vertrauenswürdiges digitales Langzeitarchiv bedient sich transparenter Technologien und Methoden zur Speicherung, Migration und Bereitstellung von Dokumenten und Metadaten. Daher arbeitet die AG Elektronisches Publizieren stetig daran, eine umfassende Dokumentation des edoc-Servers mit sämtlichen angewendeten Verfahren vorzuhalten.

Konkrete Ansätze

Für den edoc-Server der Humboldt-Universität werden mehrere Strategien verfolgt. Den Ausgangspunkt dieser Überlegungen bildet die Tatsache, dass digitalen Publikationen von unterschiedlichen Seiten betrachtet werden können:

1. Zum einen sind digitale Dokumente nichts weiter als Bitströme, die es zu erhalten und zu sichern gilt.
2. Zum anderen sind digitale Dokumente auch kodierte Informationseinheiten, die in unterschiedlichen Dateiformaten vorliegen können.
3. Zum dritten sind digitale Dokumente auch immer Informationsobjekte, die sehr komplex sein können, z. B. dadurch, dass sie aus mehreren Teilobjekten und Medienelementen bestehen können.

Backup/ Replikation und Bitstream-Preservation

Die Grundlage für alle weiteren Maßnahmen bildet daher das Konzept zum Backup und zur Replikation der Bitströme. Mit Hilfe eines SAN (Storage Area Networks) zur redundanten Speicherung mehrerer Kopien der Daten an unterschiedlichen Standorten sowie dem täglichen Backup und der täglichen Kopie der Daten wird die Existenz des Bitstroms gesichert, siehe dazu die ausführlichere Beschreibung in [5].

Zusätzlich dazu betreibt die Arbeitsgruppe Elektronisches Publizieren auch einen LOCKSS-Server und beabsichtigt, einen ausgewählten Teil des Inhaltes des edoc-Servers dadurch in weltweite Netze zu replizieren. Der Einsatz dieser Technologie ist vor allem für die über den edoc-Server veröffentlichten e-Journals geplant.

LOCKSS, *Lots Of Copies Keep Stuff Save*, wurde ursprünglich in einem Projekt der Universität Stanford entwickelt, vgl. [6]. Es ist ein so genanntes Peer-to-Peer-Netzwerk, mit dessen Hilfe Daten an mehreren Standorten verteilt und beobachtet werden. Bei Bedarf werden kaputte Bitströme durch einen Replikationsalgorithmus repariert. Ziel ist es, die Unversehrtheit von durch Bibliotheken abonnierten E-Journals auf der Bitebene zu garantieren. Dabei stellt das eigentliche LOCKSS-System die technische Infrastruktur zur Verfügung, um einen Datenverlust auf physikalischer Ebene zu verhindern. Ein Datenverlust tritt zum Beispiel ein, wenn E-Journals gestrichen werden, wenn Bibliotheken ihre Subskrip-tionen beenden müssen, sobald ein E-Journal oder ein Verlag an einen neuen Rechteinhaber verkauft wird, Webseiten von Verlagen nicht mehr erreichbar sind oder auch wenn Open-Access-Literatur aus dem WWW verschwindet. Die LOCKSS-Technologie setzt auf „Low-Tech“, d. h. es wird eine Open-Source-Software (besonderes OpenBSD, seit 1998), deren komplette Dokumentation frei zugänglich ist, genutzt. Die LOCKSS-Teilnehmer müssen eine LOCKSS-Box aufsetzen. Über das Peer-to-Peer-Netzwerk werden im Sechswochen-Takt neue Versionen der LOCKSS-Software automatisch eingespielt. Des Weiteren werden durch die Stanford-Universität automatische Tools zum Nutzen der Software (z. B. zur Erstellung von Plug-Ins) bereitgestellt. Die LOCKSS-Box entspricht einem Computer, auf dem die LOCKSS-Software läuft. Sie ist über ein Web-User-Interface konfigurier- und administrierbar und sammelt die Inhalte von Verlagsseiten, via crawl über http. Dabei können alle Formate: HTML, PDF, JPEG, TIF, Audio, Video, Excel, Java genutzt werden. Gesammelt wird die Präsentationsform der Inhalte. Eine LOCKSS-Box arbeitet als Web-Proxy, um den eigenen Inhalt anzubieten und gleicht sich über einen Peer-to-Peer-Prozess mit den anderen LOCKSS-Boxen ab. Dabei werden notfalls beschädigte Daten durch den speziellen LOCKSS-Algorithmus repariert.

Erhaltung der Benutzbarkeit und Lesbarkeit von Dokumenten – Erhaltung auf der logischen Ebene – der Ebene der Dateiformate

Dateiformate sind wohl der Faktor bei der digitalen Langzeitarchivierung, der am prägnantesten dem technischen Wandel und der Obsoletheit unterworfen sind. Die Anzahl der heute existierenden Dateiformate ist kaum abzuschätzen, da für jede erdenkliche Anwendung bzw. jedes mögliche Anwendungsprogramm eigene Formate entwickelt werden können. Für die Langzeitarchivierung wird als archivierbare Einheit das digitale Dokument angesehen. An dieser Stelle wird in Anlehnung an [7] als Dokument eine geordnete und abgeschlossene Einheit von Informationseinheiten verstanden, die in einer festen Struktur organisiert und mit Formatierungsregeln assoziiert ist, die es sowohl den Produzenten als auch Rezipienten gestattet, es zu „lesen“.

Dieses kann aus einer oder mehreren Dateien bestehen. Dabei versteht man unter einer Datei eine geordnete und abgeschlossene Sammlung von Informationseinheiten, die bei der Speicherung als Einheit betrachtet wird. Auf heutigen Systemen ist dies üblicherweise eine beliebige Anzahl eindimensional adressierter Bytes. Das Dateiformat bezeichnet alle Konventionen der inneren Struktur und Anordnung der Bytes, die die Form der Abspeicherung von Computerdateien bestimmen.

Für den edoc-Server existiert die folgende Strategie zum Umgang mit Dateiformaten: Angestrebt wird bei den meisten Publikationsvorhaben die Abspeicherung der Dokumente in einem

archivierungsfähigen Format. Das bedeutet vor allem, dass es bestimmte Eigenschaften besitzt: so muss das Format z. B. dokumentiert und offen gelegt sein, es sollte durch ein internationales, herstellerunabhängiges Gremium normiert bzw. standardisiert worden und auf verschiedenen Betriebssystemplattformen nutzbar sein. Zudem sollte es eine gute Verbreitung haben, damit die Existenz vielfältiger Softwarewerkzeuge zu dessen Verarbeitung gesichert ist. Es sollte eine Dokumentation der Revisionen bzw. eine Rückwärtskompatibilität gegeben sein. Einen wichtigen Aspekt bildet dabei auch die Tatsache, dass das Format in der Lage sein muss, die Struktur und Abfolge der Informationen so zu erfassen, dass diese extrahierbar sind (weitere Aspekte siehe [8]).

Da Markup-Sprachen aus heutiger Sicht diesen Anforderungen am ehesten gerecht werden, wurde die Nutzung XML als Strategie für die Langzeitarchivierung des edoc-Servers gewählt. Für die Publikationskategorien Dissertationen, Habilitationsschriften sowie Diplom- und Masterarbeiten kann diese Strategie zu ca. 85 Prozent umgesetzt werden. Diese Dokumente werden in der Regel von der AG Elektronisches Publizieren (EPUB-Team) nach XML konvertiert und zusätzlich zum digitalen Druckexemplar im PDF-Format gespeichert. Sie werden über ein XSLT-Stylesheet formatiert und so auf dem edoc-Server visualisiert.

Der Sinn der Nutzung von XML als Archivformat besteht darin, spätere Migrationen in neue Anzeigeformate einfacher, automatisiert und somit kostengünstiger vornehmen zu können. Die Nutzung von XML zur langfristigen Speicherung von Dokumenten hat sich in den letzten Jahren als geeignete Strategie für digitale Langzeitarchive herausgebildet und wird auch in größeren, renommierten Projekten, u. a. in PORTICO [9], eingesetzt.

Nicht alle Dokumente des edoc-Servers können nach XML konvertiert werden. Dies hat verschiedene Ursachen:

1. Die Anlieferung der Dokumente durch Autoren erfolgt bereits in einem Format, welches sich nicht ohne Weiteres nach XML konvertieren lässt. Dies ist zum Beispiel bei den Open-Access-Publikationen im edoc-Bereich *Pre- und Postprints* der Fall.
2. Das EPUB-Team verfügt bisher über keine ausgereifte und effiziente Technologie der Umwandlung. So können z. B. LateX-Dokumente nur teilweise übertragen werden. Ein Problem hierbei stellt die strukturierte und korrekte Überführung mathematischer Formeln nach XML dar.
3. Es handelt sich um einen multimedialen Bestandteil eines Dokumentes, wie z. B. eine Anlage in Form kleiner Video- oder Audiosequenzen oder gar um einen kompletten Film.

Bei Multimediainhalten wird die Strategie verfolgt, daß zum aktuellen Zeitpunkt möglichst Formate eingesetzt werden, die einen Standard im Sinne einer Norm darstellen und vor allem keine Komprimierungsalgorithmen beinhalten. Sollte zu einem zukünftigen Zeitpunkt das Format nicht mehr unmittelbar durch einen Nutzer lesbar sein, hofft man, dass es einen geeigneten Emulator für dieses Format geben wird.

So besteht die Hauptstrategie des edoc-Servers bezüglich der Langzeitarchivierung darin, zum einen die Existenz der Bitströme zu sichern und zum anderen die Vielfalt der Formate auf dem edoc-Server auf diejenigen einzuschränken, die am besten für eine Langzeitarchivierung geeignet sind. Für diese können dann auch geeignete Garantien gegeben werden. So heißt es in den Leitlinien des edoc-Servers: „Bei Verwendung des Formates SGML/XML wird eine Archivierungsgarantie von 50 Jahren gegeben. Die Archivierungsdauer anderer Formate hängt von der Verfügbarkeit des Formates, der Betrachtungssoftware sowie den Konvertierungsmöglichkeiten ab."

Als zweites Dateiformat wird auf die Nutzung von PDF/A orientiert. PDF/A ist ein ISO-Standard (ISO 19005) für langzeitarchivierbares PDF. Er schreibt detailliert vor, welche Kodierungen in PDF-Dokumenten erlaubt sind und welche nicht. In der Norm sind zwei Konformitätsebenen spezifiziert: PDF/A-1a - Level A conformance: Das Dokument muss eindeutig visuell reproduzierbar sein, Text muss nach Unicode abbildbar sein und es muss eine inhaltliche Strukturierung des Dokuments geben. Bei der PDF/A-1b - Level B conformance genügt die eindeutige visuelle Reproduzierbarkeit des Dokumentes. Verboten sind in PDF/A Referenzierungen auf externe Objekte und Ressourcen, da diese zukünftig nicht mehr zugänglich sein könnten. Um Letzteres zu verhindern, müssen in

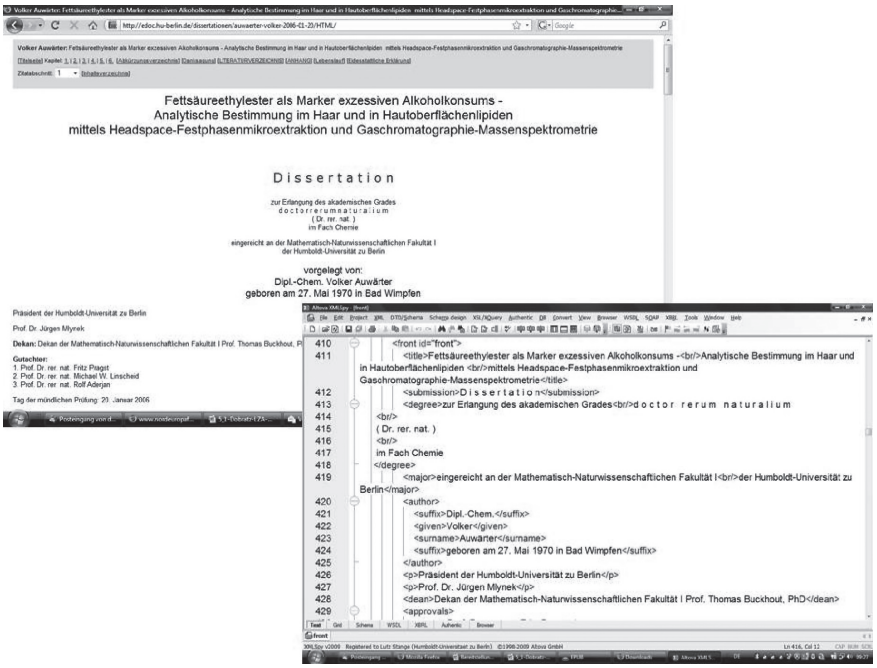

Abb. 1: Ausschnitt aus einem XML-codierten Dokument des edoc-Servers und dessen formatierte und mittels XSLT erzeugte Ansicht

einer PDF/A-1-Datei alle benutzten Schriftarten, alle Bilder, die Kennzeichnung als PDF/A-1 durch Metadaten im Extensible Metadata Platform-Format (XMP) vorhanden sein. Farben müssen, ähnlich wie in PDF/X, ausreichend definiert sein, um eine eindeutige Farbdarstellung zu garantieren. Das entsprechende Quellprofile oder ein „Output Intent" muss eingebettet werden. Es ist u. a. untersagt, alternative Bilder (z. B. niedrigaufgelöste Varianten für die Bildschirmausgabe und hochaufgelöste Varianten für den Druck) zu integrieren. PDF/A gestattet keine Verschlüsselungen oder das Sperren von Funktionen der Datei wie z. B. das Drucken und das Kopieren von Daten aus der PDF-Datei heraus. Weiterhin ist die Einbettung von Programmiercode, z. B. JavaScript, verboten, da dessen Ausführung den Inhalt oder die Darstellung des Dokumentes verändern kann. Audio- oder Videodaten dürfen nicht in PDF/A-Dokumente integriert sein. Allerdings unterstützt PDF/A die Einbettung von digitalen Signaturen, vgl. [10]. Durch diese und andere Vorschriften soll eine langfristige Benutzbarkeit der Dokumente garantiert werden. PDF/A ist unabhängig von jedweder Anwendungssoftware oder Betriebssystemen.

Authentizität und Integrität digitaler Dokumente

Im Prinzip ist es für alle edoc-Publikationen wichtig, dauerhaft und rechtlich relevant nachweisen zu können, dass die digitalen Dokumente vollständig und unverändert auf dem edoc-Server verfügbar sind. Dies nennt man Sicherung der Integrität der Dokumente. Zum anderen muss für einen Teil der Dokumente nachweisbar sein, dass der Ersteller und der Veröffentlichungszeitpunkt nachweisbar sind, die auf dem edoc-Server angegeben werden. Dies nennt man Sicherung der Authentizität.

Mit der Erstellung entsprechender Hashwerte sowie dem Anbringen digitaler Signaturen und Zeitstempel an diese können für die edoc-Dokumente, besonders für die Qualifikationsarbeiten, rechtssicher deren Integrität und Authentizität gesichert werden. Die Details zu dem Verfahren sind im Artikel [10] nachzulesen.

Die Nutzung digitaler Hashwerte und Signaturen allein verhindert natürlich nicht, dass Manipulationen an den Dokumenten vorgenommen werden. Hierzu sind weitere organisatorische Sicherheitsmaßnahmen wie Zutritts- und Zugangskontrolle zu den Serverräumen sowie die Zugriffskontrolle durch Vergabe von Benutzerrechten auf das System des edoc-Servers und die darin befindlichen Dokumente erforderlich.

Hashwerte und digitale Signaturen können allerdings dazu dienen, unautorisierte Änderungen an den Dokumenten festzustellen.

Metadaten und dauerhafte Adressierbarkeit

Für ein langfristiges Management der im digitalen Archiv befindlichen Dokumente benötigt man geeignete Metadaten. Bisher werden für den edoc-Server vorrangig bibliographische und organisatorische Metadaten erfasst. Technische Metadaten werden derzeit nur ausschnittsweise gespeichert. Hier besteht die Notwendigkeit, die Menge der erfassten technischen Metadaten so zu erweitern, dass spezifischere Informationen zu den Dokumentformaten und der Erstellungssoftware in Anlehnung an das METS-Metadatenmodell, kombiniert mit dem Metadatenmodell der PREMIS-Arbeitsgruppe oder alternativ dem LMER-Metadatenmodell, mit aufgenommen werden. Ein Konzept zur Speicherung der Informationspakete mit den zu jeder Version bzw. jedem Format gehörenden Metadaten wurde bereits erarbeitet und befindet sich derzeit in der Umsetzungsphase, siehe dazu [11]. Dabei wird beim edoc-Server der Humboldt-Universität zu Berlin ein besonderer Wert auf die Einbindung digitaler Signaturen in ein Metadatenkonzept gelegt.

Ein für die Langzeitarchivierung entscheidender Parameter ist die Vergabe konsistenter und dauerhafter Adressen, um die Wiederauffindbarkeit der Dokumente zu gewährleisten. Für alle Dokumente auf dem edoc-Server der Humboldt-Universität werden persistente Identifikatoren in Form der URN (Uniform Ressource Name) vergeben. Dabei wird auf das Konzept der National Bibliographic Number (nbn) der Deutschen Nationalbibliothek zurückgegriffen, vgl. [12].

So wird zum Beispiel für das Dokument mit der URL *http://edoc.hu-berlin.de/series/nestor-materialien/2008-8/PDF/8.pdf*, folgende URN vergeben: *urn:nbn:de:0008-20080218o2*. Sowohl der edoc-Server als auch der Resolverdienst bei der Deutschen Nationalbibliothek führen eine Datenbank, in der beide Zeichenketten einander zugeordnet werden. Ändert sich zu einem späteren Zeitpunkt der physische Speicherort des Dokumentes z. B. auf *future1.cms.hu-berlin.de* wird in der Resolverdatenbank die der URN zugeordnete URL aktualisiert und der Nutzer kann trotzdem weiter mit der alten URN auf das nun an anderer Stelle gespeicherte Dokument referenzieren.

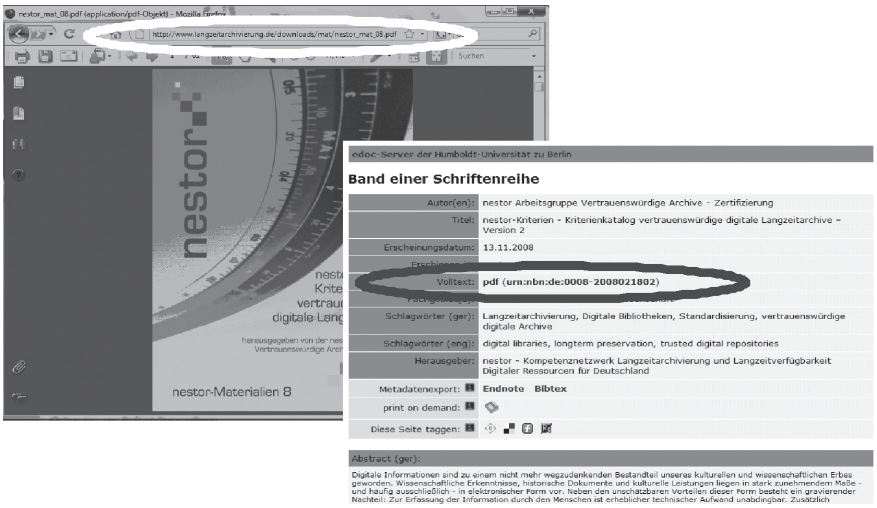

Abb. 2: Beispiel für eine URN in Form der nbn auf dem edoc-Server (der Aufruf des Dokumentes erfolgt über: http://nbn-resolving.de/urn:nbn:de:0008-20080218o2)

Pflichtabgabegesetz

Langzeitarchivierung ist eine nationale und Institutionen übergreifende Aufgabe. Dies ist ein Grund für das seit Juni 2006 existierende „Gesetz über die Deutsche Nationalbibliothek“. In diesem ist auch die Pflichtabgabe elektronischer Dokumente, wie sie z. B. hier über den edoc-Server publiziert werden, geregelt. Eine automatische Bereitstellung der Informationen zu den neuen, auf dem edoc-Server veröffentlichten Dokumenten erfolgt über die Bereitstellung der URN-Informationen über die OAI-Schnittstelle des edoc-Servers. Diese kann von der Deutschen Nationalbibliothek abgefragt und die Dokumente und deren Metadaten können so direkt abgerufen werden. Zurzeit ist ein vollautomatisiertes Melde- und Abgabeverfahren für Dokumente an die DNB zwar noch nicht im Produktivbetrieb, es wird aber davon ausgegangen, dass dieses in naher Zukunft der Fall sein wird. Dann kann auch dafür garantiert werden, dass alle Dokumente des edoc-Servers zusätzlich in der DNB archiviert werden.

Für die digitalen Dissertationen wird ein derartiges Verfahren bereits seit einigen Jahren erfolgreich von der DNB durchgeführt. Die Dokumente sind dann noch einmal über den Deposit-Server in der DNB verfügbar.

Fazit

Die Humboldt-Universität zu Berlin ist mit dem edoc-Server auf einem guten Weg, für digital publizierte Dokumente ein geeignetes Langzeitarchiv bereitzustellen. Obwohl das bisherige System (noch) nicht den Anforderungen der „Vertrauenswürdigen digitalen Langzeitarchive„, vgl. [8] entspricht und in jedem Detail dem OAIS-Modell [13] folgt, sind die Betreiber zuversichtlich, die Dokumente auch zukünftigen Generationen von Wissenschaftlern und Studierenden in einer benutzbaren Form anbieten zu können. Die weitere Arbeit an den technischen Details wird permanent fortgeführt. Im aktuellen Fokus steht dabei die Umsetzung einer geeigneten Metadatenstrategie, vgl. [11].

Literatur

[1] Rothenberg, Jeff: *Digital Information Lasts Forever – Or Five Years, Whichever Comes First*. 2. Oktober 2001, Vortrag, http://www.amibusiness.com/dps/rothenberg-arma.pdf (18.03.2009)

[2] Arbeitsgruppe Elektronisches Publizieren: *Dokumenten- und Publikationsserver der Humboldt-Universität zu Berlin – Leitlinien –, Humboldt-Universität zu Berlin*. 2001, http://edoc.hu-berlin.de/e_info/leitlinien.php (17.03.2009)

[3] Schwens, Ute; Liegmann, Hans: *Die digitale Welt – eine ständige Herausforderung*. Kuhlen, Rainer; Seeger, Thomas und Strauch, Dietmar, Handbuch zur Einführung in die Informationswissenschaft und -praxis, 5. völlig neu gefasste Ausgabe, Auflage, München, (2004)

[4] nestor - Materialien 8: *nestor – Kompetenznetzwerk Langzeitarchivierung / Arbeitsgruppe Vertrauenswürdige Archive – Zertifizierung: nestor-Kriterien*. Kriterienkatalog vertrauenswürdige digitale Langzeitarchive, Version 2, 2008, Frankfurt am Main : nestor c/o Deutsche Nationalbibliothek, urn:nbn:de:0008-2008021802 (18.03.2009)

[5] Fromm, Niels: *Backup-Strategie für den Dokumentenserver, cms-journal 32, 2009, Humboldt-Universität zu Berlin, (16.03.3009)*

[6] Reich, Vicky; David S. Rosenthal; LOCKSS: *A Permanent Web Publishing and Access System*. D-Lib Magazine, Volume 7 Number 6, June 2001, DOI: 10.1045/june2001-reich, http://www.dlib.org/dlib/june01/reich/06reich.html (20.02.2009)

[7] Pédauque, Roger T.: *Document: Form, Sign and Medium, As Reformulated for Electronic Documents*. (2003) URL: http://archivesic.ccsd.cnrs.fr/docs/00/06/22/28/PDF/sic_00000594.pdf, (18.03.2009)

[8] Stanescu, A.: *Assessing the Durability of Formats in a Digital Preservation Environment*. D-Lib Magazine (Band 10), Nr. 11., 2005, URL: http://www.dlib.org/dlib/november04/stanescu/11stanescu.html (17.03.200)

[9] PORTICO: http://www.portico.org/ (17.03.2009)

[10] Fromm, Niels: *Signatur und Zeitstempel zur Wahrung von Authentizität und Integrität*. cms-journal 32, 2009, Humboldt-Universität zu Berlin, (16.03.3009)

[11] Fromm, Niels: *Einsatz elektronischer Signaturen auf dem Dokumentenserver der Humboldt-Universität zu Berlin*. Diplomarbeit, Humboldt-Universität zu Berlin, Institut für Informatik, 2009

[12] Schöning-Walter, Christa: *Der Uniform Resource Name (URN)*. In: nestor-Handbuch, Version 1.2 (Juni 2008), Kapitel 13.2.1, http://nestor.sub.uni-goettingen.de/handbuch/artikel/nestor_handbuch_artikel_156.pdf (18.03.2009)

[13] CCSDS (Consultative Committee for Space Data Systems): *Reference Model for an Open Archival Information System (OAIS)*. Blue Book ISO 14721: 2003 (Band Issue 1). URL: http://www.ccsds.org/docu/dscgi/ds.py/Get/File-143/650x0b1.pdf (18.03.2009)